Title:

Subtitle

©Patrick Hall 20XX

May 26, 2024

Abstract

1 Introduction

The National Institute of Standards and Technology Artificial Intelligence (AI) Risk Management Framework (RMF).[1]

2 Generative AI Governance

3 Generative AI Inventories

4 Generative AI Risk Tiers

5 Generative AI Risk Measurement

6 Generative AI Risk Management

Conclusion

Acknowledgments

Thank you to Bernie Siskin and Nick Schmidt of BLDS and Eric Sublett of Relman Colfax for formative discussions relating to GAI risk tiering.

Abbreviations

- AI: Artificial Intelligence
- AI RMF: Artificial Intelligence Risk Management Framework
- GAI: Generative AI
- RMF: Risk Management Framework

[1] NIST AI. Artificial Intelligence Risk Management Framework (AI RMF 1.0). 2023.

[2] Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*, 2023.

[3] Rishi Bommasani, Percy Liang, and Tony Lee. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023.

[4] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

[5] Adrian de Wynter, Xun Wang, Alex Sokolov, Qilong Gu, and Si-Qing Chen. An evaluation on large language model outputs: Discourse and memorization. *Natural Language Processing Journal*, 4:100024, 2023.

[6] Jeremy Dohmann. Blazingly fast llm evaluation for in-context learning. https://www.databricks.com/blog/llm-evaluation-for-icl. Last accessed: May 24, 2024.

[7] Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? *arXiv:2402.07841*, 2024.

[8] Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

[9] Hugging Face. Evaluation. https://huggingface.co/docs/evaluate/index. Last accessed: May 24, 2024.

[10] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*, 2023.

[11] Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv:2204.08582*, 2022.

[12] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023.

[13] Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation. In *The Twelfth International Conference on Learning Representations*, 2023.

[14] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36, 2024.

[15] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

[16] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box llms automatically. *arXiv preprint arXiv:2312.02119*, 2023.

[17] Julien Piet, Chawin Sitawarin, Vivian Fang, Norman Mu, and David Wagner. Mark my words: Analyzing and evaluating language model watermarks. *arXiv preprint arXiv:2312.00273*, 2023.

[18] Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, Markus Pauly, et al. The self-perception and political biases of chatgpt. *Human Behavior and Emerging Technologies*, 2024.

[19] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*, 2023.

[20] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

[21] Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. *arXiv preprint arXiv:2205.09209*, 2022.

[22] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[23] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*, 2023.

[24] Bertie Vidgen, Adarsh Agrawal, Ahmed M Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Borhane Blili-Hamelin, et al. Introducing v0. 5 of the ai safety benchmark from mlcommons. *arXiv preprint arXiv:2404.12241*, 2024.

[25] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Advances in Neural Information Processing Systems*, 36, 2024.

[26] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*, 2023.

[27] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

Appendix A: Example Generative AI–Trustworthy Characteristic Crosswalk

6.1 A.1: Trustworthy Characteristic to Generative AI Risk Crosswalk

Table 1: Trustworthy Characteristic to Generative AI Risk Crosswalk.

Accountable and Transparent	Explainable and Interpretable	Fair with Harmful Bias Managed	Privacy Enhanced	Safe	Secure and Resilient	Valid and Reliable
Data Privacy Environmental Human-AI Configuration Information Integrity Intellectual Property Value Chain and Component Integration	Human-AI Configuration Value Chain and Component Integration	Confabulation Environmental Human-AI Configuration Intellectual Property Obscene, Degrading, and/or Abusive Content Toxicity, Bias, and Homogenization Value Chain and Component Integration	Data Privacy Human-AI Configuration Information Security Intellectual Property Value Chain and Component Integration	CBRN Information Confabulation Dangerous or Violent Recommendations Data Privacy Environmental Human-AI Configuration Information Integrity Information Security Obscene, Degrading, and/or Abusive Content Value Chain and Component Integration	Dangerous or Violent Recommendations Data Privacy Human-AI Configuration Information Security Value Chain and Component Integration	Confabulation Human-AI Configuration Information Integrity Information Security Toxicity, Bias, and Homogenization Value Chain and Component Integration

6.2 A.2: Generative AI Risk to Trustworthy Characteristic Crosswalk

Table 2: Generative AI Risk to Trustworthy Characteristic Crosswalk.

CBRN Information	Confabulation	Dangerous or Violent Recommendations	Data Privacy
Safe	Fair with Harmful Bias Managed Safe Valid and Reliable	Safe Secure and Resilient	Accountable and Transparent Privacy Enhanced Safe Secure and Resilient

Environmental	Human-AI Configuration	Information Integrity	Information Security
Accountable and Transparent Fair with Harmful Bias Managed Safe	Accountable and Transparent Explainable and Interpretable Fair with Harmful Bias Managed Privacy Enhanced Safe Secure and Resilient Valid and Reliable	Accountable and Transparent Safe Valid and Reliable	Privacy Enhanced Safe Secure and Resilient Valid and Reliable

Intellectual Property	Obscene, Degrading, and/or Abusive Content	Toxicity, Bias, and Homogenization	Value Chain and Component Integration
Accountable and Transparent Fair with Harmful Bias Managed Privacy Enhanced	Fair with Harmful Bias Managed Safe	Fair with Harmful Bias Managed Valid and Reliable	Accountable and Transparent Explainable and Interpretable Fair with Harmful Bias Managed Privacy Enhanced Safe Secure and Resilient Valid and Reliable

Appendix B: Example Risk Tiers for Generative AI

Appendix C: List of Publicly Available Model Testing Suites (“Evals”)

C.1: Publicly Available Model Testing Suites (“Evals”) by Trustworthy Characteristic

Table 3: Publicly Available Model Testing Suites (“Evals”) by Trustworthy Characteristic.

Accountable and Transparent
An Evaluation on Large Language Model Outputs: Discourse and Memorization (see Appendix B)[5]
Big-bench: Truthfulness [22]
DecodingTrust: Machine Ethics [25]
Evaluation Harness: ETHICS [12]
HELM: Copyright [3]
Mark My Words [17]

Fair with Harmful Bias Managed
BELEBELE [2]
Big-bench: Low-resource language, Non-English, Translation
Big-bench: Social bias, Racial bias, Gender bias, Religious bias
Big-bench: Toxicity
DecodingTrust: Fairness
DecodingTrust: Stereotype Bias
DecodingTrust: Toxicity
C-Eval (Chinese evaluation suite) [14]
Evaluation Harness: CrowS-Pairs
Evaluation Harness: ToxiGen
Finding New Biases in Language Models with a Holistic Descriptor Dataset [21]
From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models [10]
HELM: Bias
HELM: Toxicity
MT-bench [27]
The Self-Perception and Political Biases of ChatGPT [18]
Towards Measuring the Representation of Subjective Global Opinions in Language Models [8]

Privacy Enhanced
HELM: Copyright
llmprivacy [23]
mimir [7]

Safe
Big-bench: Convince Me
Big-bench: Truthfulness
HELM: Reiteration, Wedging
Mark My Words
MLCommons [24]
The WMDP Benchmark [15]

Publicly Available Model Testing Suites (“Evals”) by Trustworthy Characteristic (continued).

Secure and Resilient
Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation [13]
DecodingTrust: Adversarial Robustness, Robustness Against Adversarial Demonstrations
detect-pretrain-code [20]
In-The-Wild Jailbreak Prompts on LLMs [19]
JailbreakingLLMs [4]
llmprivacy
mimir
TAP: A Query-Efficient Method for Jailbreaking Black-Box LLMs [16]

Valid and Reliable
Big-bench: Algorithms, Logical reasoning, Implicit reasoning, Mathematics, Arithmetic, Algebra, Mathematical proof, Fallacy, Negation, Computer code, Probabilistic reasoning, Social reasoning, Analogical reasoning, Multi-step, Understanding the World
Big-bench: Analytic entailment, Formal fallacies and syllogisms with negation, Entailed polarity
Big-bench: Context Free Question Answering
Big-bench: Contextual question answering, Reading comprehension, Question generation
Big-bench: Morphology, Grammar, Syntax
Big-bench: Out-of-Distribution
Big-bench: Paraphrase
Big-bench: Sufficient information
Big-bench: Summarization
DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations
Eval Gauntlet: Reading comprehension [6]
Eval Gauntlet: Commonsense reasoning, Symbolic problem solving, Programming
Eval Gauntlet: Language Understanding
Eval Gauntlet: World Knowledge
Evaluation Harness: BLiMP
Evaluation Harness: CoQA, ARC
Evaluation Harness: GLUE
Evaluation Harness: HellaSwag, OpenBookQA, TruthfulQA
Evaluation Harness: MuTual
Evaluation Harness: PIQA, PROST, MC-TACO, MathQA, LogiQA, DROP
FLASK: Logical correctness, Logical robustness, Logical efficiency, Comprehension, Completeness [26]
FLASK: Readability, Conciseness, Insightfulness
HELM: Knowledge
HELM: Language
HELM: Text classification
HELM: Question answering
HELM: Reasoning
HELM: Robustness to contrast sets
HELM: Summarization
Hugging Face: Fill-mask, Text generation [9]
Hugging Face: Question answering
Hugging Face: Summarization
Hugging Face: Text classification, Token classification, Zero-shot classification
MASSIVE [11]
MT-bench

C.2: Publicly Available Model Testing Suites (“Evals”) by Generative AI Risk

Table 4: Publicly Available Model Testing Suites (“Evals”) by Generative AI Risk.

CBRN Information
Big-bench: Convince Me
Big-bench: Truthfulness
HELM: Reiteration, Wedging
MLCommons
The WMDP Benchmark

Confabulation
BELEBELE
Big-bench: Algorithms, Logical reasoning, Implicit reasoning, Mathematics, Arithmetic, Algebra, Mathematical proof, Fallacy, Negation, Computer code, Probabilistic reasoning, Social reasoning, Analogical reasoning, Multi-step, Understanding the World
Big-bench: Analytic entailment, Formal fallacies and syllogisms with negation, Entailed polarity
Big-bench: Context Free Question Answering
Big-bench: Contextual question answering, Reading comprehension, Question generation
Big-bench: Convince Me
Big-bench: Low-resource language, Non-English, Translation
Big-bench: Morphology, Grammar, Syntax
Big-bench: Out-of-Distribution
Big-bench: Paraphrase
Big-bench: Sufficient information
Big-bench: Summarization
Big-bench: Truthfulness
C-Eval (Chinese evaluation suite)
DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations
Eval Gauntlet Reading comprehension
Eval Gauntlet: Commonsense reasoning, Symbolic problem solving, Programming
Eval Gauntlet: Language Understanding
Eval Gauntlet: World Knowledge
Evaluation Harness: BLiMP
Evaluation Harness: CoQA, ARC
Evaluation Harness: GLUE
Evaluation Harness: HellaSwag, OpenBookQA, TruthfulQA
Evaluation Harness: MuTual
Evaluation Harness: PIQA, PROST, MC-TACO, MathQA, LogiQA, DROP
FLASK: Logical correctness, Logical robustness, Logical efficiency, Comprehension, Completeness
FLASK: Readability, Conciseness, Insightfulness
Finding New Biases in Language Models with a Holistic Descriptor Dataset
HELM: Knowledge
HELM: Language
HELM: Language (Twitter AAE)
HELM: Question answering
HELM: Reasoning
HELM: Reiteration, Wedging
HELM: Robustness to contrast sets
HELM: Summarization
HELM: Text classification
Hugging Face: Fill-mask, Text generation
Hugging Face: Question answering
Hugging Face: Summarization
Hugging Face: Text classification, Token classification, Zero-shot classification
MASSIVE
MLCommons
MT-bench

Publicly Available Model Testing Suites (“Evals”) by Generative AI Risk (continued).

Dangerous or Violent Recommendations
Big-bench: Convince Me
Big-bench: Toxicity
DecodingTrust: Adversarial Robustness, Robustness Against Adversarial Demonstrations
DecodingTrust: Machine Ethics
DecodingTrust: Toxicity
Evaluation Harness: ToxiGen
HELM: Reiteration, Wedging
HELM: Toxicity
MLCommons

Data Privacy
An Evaluation on Large Language Model Outputs: Discourse and Memorization (with human scoring, see Appendix B)
Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation
DecodingTrust: Machine Ethics
Evaluation Harness: ETHICS
HELM: Copyright
In-The-Wild Jailbreak Prompts on LLMs
JailbreakingLLMs
MLCommons
Mark My Words
TAP: A Query-Efficient Method for Jailbreaking Black-Box LLMs
detect-pretrain-code
llmprivacy
mimir

Environmental
HELM: Efficiency

Information Integrity
Big-bench: Analytic entailment, Formal fallacies and syllogisms with negation, Entailed polarity
Big-bench: Convince Me
Big-bench: Paraphrase
Big-bench: Sufficient information
Big-bench: Summarization
Big-bench: Truthfulness
DecodingTrust: Machine Ethics
DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations
Eval Gauntlet: Language Understanding
Eval Gauntlet: World Knowledge
Evaluation Harness: CoQA, ARC
Evaluation Harness: ETHICS
Evaluation Harness: GLUE
Evaluation Harness: HellaSwag, OpenBookQA, TruthfulQA
Evaluation Harness: MuTual
Evaluation Harness: PIQA, PROST, MC-TACO, MathQA, LogiQA, DROP
FLASK: Logical correctness, Logical robustness, Logical efficiency, Comprehension, Completeness
FLASK: Readability, Conciseness, Insightfulness
HELM: Knowledge
HELM: Language
HELM: Question answering
HELM: Reasoning
HELM: Reiteration, Wedging
HELM: Robustness to contrast sets
HELM: Summarization
HELM: Text classification
Hugging Face: Fill-mask, Text generation
Hugging Face: Question answering
Hugging Face: Summarization
MLCommons
MT-bench
Mark My Words

Publicly Available Model Testing Suites (“Evals”) by Generative AI Risk (continued).

Information Security
Big-bench: Convince Me
Big-bench: Out-of-Distribution
Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation
DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations
Eval Gauntlet: Commonsense reasoning, Symbolic problem solving, Programming
HELM: Copyright
In-The-Wild Jailbreak Prompts on LLMs
JailbreakingLLMs
Mark My Words
TAP: A Query-Efficient Method for Jailbreaking Black-Box LLMs
detect-pretrain-code
llmprivacy
mimir

Intellectual Property
An Evaluation on Large Language Model Outputs: Discourse and Memorization (with human scoring, see Appendix B)
HELM: Copyright
Mark My Words
llmprivacy
mimir

Obscene, Degrading, and/or Abusive Content
Big-bench: Social bias, Racial bias, Gender bias, Religious bias
Big-bench: Toxicity
DecodingTrust: Fairness
DecodingTrust: Stereotype Bias
DecodingTrust: Toxicity
Evaluation Harness: CrowS-Pairs
Evaluation Harness: ToxiGen
HELM: Bias
HELM: Toxicity

Toxicity, Bias, and Homogenization
BELEBELE
Big-bench: Low-resource language, Non-English, Translation
Big-bench: Out-of-Distribution
Big-bench: Social bias, Racial bias, Gender bias, Religious bias
Big-bench: Toxicity
C-Eval (Chinese evaluation suite)
DecodingTrust: Fairness
DecodingTrust: Stereotype Bias
DecodingTrust: Toxicity
Eval Gauntlet: World Knowledge
Evaluation Harness: CrowS-Pairs
Evaluation Harness: ToxiGen
Finding New Biases in Language Models with a Holistic Descriptor Dataset
From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models
HELM: Bias
HELM: Toxicity
The Self-Perception and Political Biases of ChatGPT
Towards Measuring the Representation of Subjective Global Opinions in Language Models

Appendix D: List of Common Adversarial Prompting Strategies

D.1: Common Adversarial Prompting Strategies by Trustworthy Characteristic

D.2: Common Adversarial Prompting Strategies by Generative AI Risk

Appendix E: Common Risk Controls for Generative AI

E.1: Common Risk Controls for Generative AI by Trustworthy Characteristic

E.2: Common Risk Controls for Generative AI by Generative AI Risk

Appendix F: Example Low-risk Generative AI Measurement and Management Plan

6.3 F.1: Example Low-risk Generative AI Measurement and Management Plan by Trustworthy Characteristic

6.4 F.2: Example Low-risk Generative AI Measurement and Management Plan by Generative AI Risk

Appendix G: Example Medium-risk Generative AI Measurement and Management Plan

6.5 G.1: Example Medium-risk Generative AI Measurement and Management Plan by Trustworthy Characteristic

6.6 G.2: Example Medium-risk Generative AI Measurement and Management Plan by Generative AI Risk

Appendix H: Example High-risk Generative AI Measurement and Management Plan

6.7 H.1: Example High-risk Generative AI Measurement and Management Plan by Trustworthy Characteristic

6.8 H.2: Example High-risk Generative AI Measurement and Management Plan by Generative AI Risk